Abstract (< 500 words)

The analysis of how genome information creates phenotypes at the single cell level, the fundamental unit of biology, is ***a powerful approach for understanding genome function***, and is rapidly becoming the gold standard for human genetics research predicting phenotype from genotype. The multicellular complexity of plant and animal agricultural species limits our understanding of the regulation and organization of their genome, and the expression patterns of their genes in each cell composing these species. To make the enormous promise of single-cell (SC) genomics a reality for the agricultural genome to phenome community, we need to develop Findable, Accessible, Interoperable, and Reuseable (FAIR) SC data resources and informatic tools for storing, sharing, and analyzing such data that is currently accumulating in crop and livestock research groups. We believe this *Enabling* seed grant proposal addresses topic areas #1 and 2 in the AG2PI RFP. The lack of FAIR SC data and the computational skills required for researchers to use such data currently prevents the adoption of this powerful method within the AG2PI community.

Our long-term goal is to construct a scientist-friendly data resource and analytical ecosystem to facilitate SC level genomic analysis through data storage, retrieval, re-use, visualization and comparative annotation across agricultural species. This goal will be actualized through a Working Group that will write a USDA-AFRI proposal based on the proof-of-principle (POP) studies we propose in this seed grant, which leverage the world-class resources in the Human Cell Atlas Data Coordination Platform (HCA DCP). In Aim 1, an early-career scientist will collaborate with a cross-kingdom group of researchers and data scientists, including those at the HCA DCP, to test the utility of current crop and livestock metadata standards, learn the use of tools and systems within the HCA DCP environment, and document the results and SC analytical tools available there. In Aim 2, we will develop solutions to the lack of scRNAseq data visualization at genome browsers for Ag species. We will test existing tools and develop new tools to map SC data results in an Ag species genome context, initially using published datasets and following later with output from Aim 1 analyses.

This seed grant brings together the crop and livestock communities at the very early stage of creating and analyzing SC data to ensure its effective use for G2P research. It is therefore a very opportune time to coordinate such a discussion with the HCA community who have built a world-leading SC data infrastructure. We will produce several deliverables, including a demonstration that current crop and livestock metadata can be used to ingest scRNAseq data at HCA DCP and that tools within the HCA DCP can be used to analyze ingested data for further use. We also will produce a white paper and website page summarizing tools for scRNAseq data exploration, visualization, and analysis, further increasing FAIR-ness of Ag SC data. We will produce initial tools for displaying scRNAseq data on genome browsers and querying outputs of scRNAseq analysis in a web-accessible database.

Project description (**no more than four pages**)

1. Objectives/aims

Aim 1: Using proof-of-principle, demonstrate a method to dramatically improve availability of FAIR crop and livestock single cell (SC) data/metadata and analytical workflows.
Aim 2: Test and develop prototype tools to make the output of scRNAseq analyses available for visualization on genome browsers and querying in databases for agriculturally important species.

Rationale: Genome annotation in livestock and crop species is currently focused on tissue-based analyses-a transition to SC-based work is needed.

The PI and many global colleagues have organized FAANG, a world-wide initiative to annotate the functional components of domesticated species using epigenetic analysis primarily of tissues and organs (1). Similar efforts are underway in several plant species (2). However, tissue-level annotation is less sensitive, and potentially misleading, than analysis at the single-cell (SC) level, as the "tissue-averaged" signals may not allow detection of expression patterns from rare cell populations, and different regulatory pathways functioning in different cells within a tissue may cancel gene expression signals (3). Plant and animal communities have recognized the value of SC methods (1,4,5), and we have begun publishing scRNAseq analyses on crops and livestock(6–8). However, unlike the population-based sequencing assays listed above, SC data is more complex to analyze and interpret (9), and there are few standards available to make SC data FAIR. A scientist-friendly ecosystem is needed for optimal use and re-use of such important data for the ag community. As single-cell genomics crop and livestock researchers and data scientists within the Human Cell Atlas (HCA) program, we have developed a plan that begins to address these gaps in data infrastructure (Fig. 1).

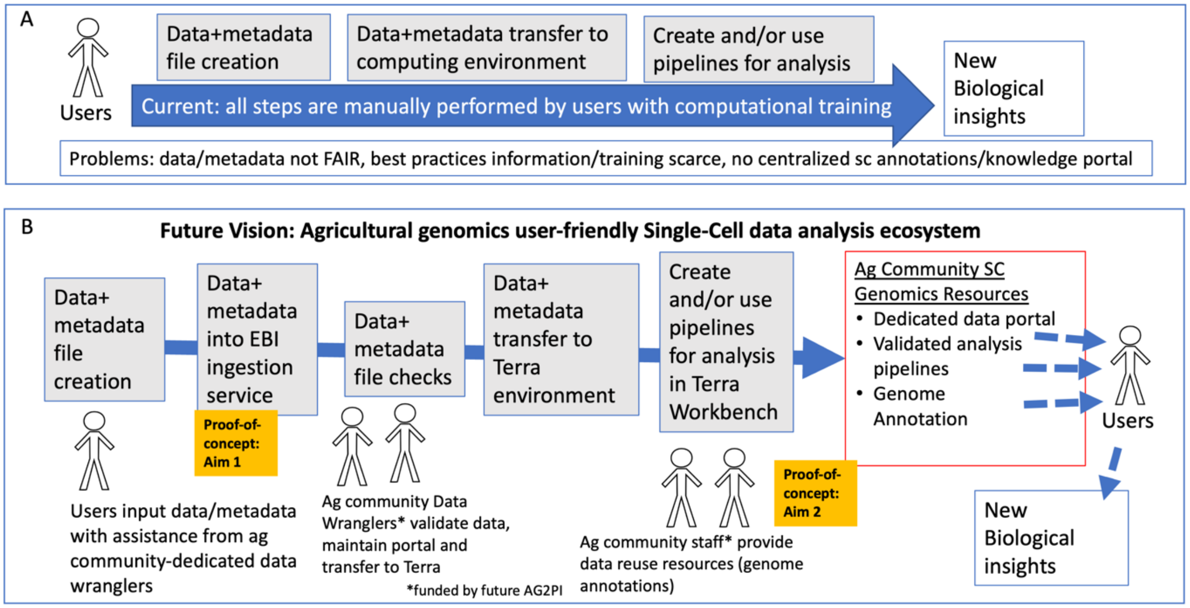
A
Users
Data+metadata file creation
Data+metadata transfer to computing environment
Create and/or use pipelines for analysis
Current: all steps are manually performed by users with computational training
New Biological insights
Problems: data/metadata not FAIR, best practices information/training scarce, no centralized sc annotations/knowledge portal
B
Future Vision: Agricultural genomics user-friendly Single-Cell data analysis ecosystem
Data+ metadata file creation
Data+ metadata into EBI ingestion service
Proof-of-concept: Aim 1
Data+ metadata file checks
Data+ metadata transfer to Terra environment
Create and/or use pipelines for analysis in Terra Workbench
Ag Community SC Genomics Resources
• Dedicated data portal
• Validated analysis pipelines
• Genome Annotation
Users
Users input data/metadata with assistance from ag community-dedicated data wranglers
Ag community Data Wranglers* validate data, maintain portal and transfer to Terra
*funded by future AG2PI
Ag community staff* provide data reuse resources (genome annotations)
Proof-of-concept: Aim 2
New Biological insights

Fig 1. (A) Current Status and (B) Future Vision for Single Cell Data analysis in Agriculture

Approach

First, we note there is a hodge-podge of data/metadata submission standards, and we propose to determine whether the world-class infrastructure of the HCA program can be readily leveraged for agricultural genomics. The HCA Data Coordination Platform (HCADCP) data browser allows a search of >20 million cells across >10,000 specimens, and displays links to metadata and matrices (https://data.humancellatlas.org/explore/projects). The latter is produced from raw submitted data files using uniform pipelines, including generating QC metrics. Once data is ingested, it is validated, and then ported to Terra, the Broad Institute scalable open access cloud platform for biomedical researchers to access Human Cell Atlas (HCA) data, run analysis and collaborate (https://data.humancellatlas.org/analyze/portals/terra).

The purpose of Aim 1 will be to determine, via a proof of principle (POP) test, if our current metadata standards in crops and livestock can be used to ingest selected 10X scRNAseq datasets (2 each from plant and animal species) into the HCA-DCP for further analysis (Figure 1). Collaborators will assist in dataset selection; Ben Cole and Marc Libault for plants, Wes Warren for animals (see letters). Co-PI Harrison will provide FAANG data standards and support to Muskan Kapoor, the ISU PhD student performing this POP data ingestion. Plant data standards for SC data are evolving; Kapoor will start with MINSEQE (https://zenodo.org/record/5706412), and co-PI Provart will assist and add plant/environmental ontology metadata when possible (see letter). We will initially use our published datasets, to allow comparison of analytical results (see below) to our manually created ones (6,7). We expect to find that modification of these standards is necessary, and EBI collaborator Tony Burdett (see letter) will provide support to Kapoor.

Further, a major bottleneck in researchers re-using SC data is the large and complex collection of tools available for SC data analysis (9), and the lack of any standardization or authentication of pipelines within the Ag community. However, excellent such infrastructure exists within the HCA community at Terra. Our collaborator, Tim Tickle, at the Broad will consult and provide support to Kapoor to test Terra workflows with the POP ingested and processed datasets (see letter). For 10X scRNAseq data, both the Optimus and Cumulus pipelines are available. Kapoor will self-train using both HCA-DCP and Terra online training modules, as well as attend on-line training events (https://support.terra.bio/hc/en-us). In addition to testing these workflows, she will write a summary of Terra usage/capabilities for the seed grant report and at www.ag2pi.org. Provart is compiling a summary of SC analytical workflows for the PCA consortium and will work with Kapoor on this component. Terra provides interactive environments including Jupyter notebooks and RStudio instances supporting Bioconductor, Seurat, Scanpy, Pegasus, and other programming libraries. Kapoor has experience analyzing scRNAseq data with Jupyter notebooks and will evaluate and document the use of these environments to analyze the POP datasets. We also anticipate using our own additional processed data to validate workflows and will summarize for the report and at www.ag2pi.org.

Second, we will develop a solution to the lack of scRNAseq data visualization for genome browsers for Ag species. In Aim 2, staff in the Chris Elsik lab (see letter) will test existing tools for visualizing scRNAseq in a genomic context, and will develop tracks to display on JBrowse (10). Read alignment and histogram tracks to show expression levels along the genome for cell types or clusters will be created by developing code to generate read alignment (BAM) files for individual cell clusters by matching BAM file barcodes with the cluster file barcodes from the scRNAseq analysis pipeline. In addition to cluster-specific read alignment tracks, we will use utilities available from the UCSC Genome Browser (11) to generate bigBed files from the original BAM, expression matrix and cluster file to create tissue-based tracks that display a histogram of cluster/cell type expression for each gene, similar to the scRNAseq tracks on the UCSC human genome browser. Cluster information and experiment metadata will be used to create a JBrowse faceted track selector that will allow users to filter tracks for viewing by tissue, unique cluster identifier, cell type (if known), NCBI and/or EMBL experiment and biosample accessions, PubMed accession and other types of metadata. In addition to making cluster/cell type tracks available for viewing on JBrowse, we will load metadata and gene expression data into an InterMine (12) database, similar to the FAANGMine, MaizeMine. and ThaleMine databases currently hosted by the Elsik and Provart labs. For loading expression data into an InterMine database, a tab-delimited file of gene expression counts per cluster will be generated using the expression matrix file and cluster file generated in the scRNAseq analysis. Initially tool testing and development will leverage existing data, and then will transition to using output generated by Kapoor on the Terra platform. Working closely with ISU, the Elsik lab will work to streamline the uptake of pipeline outputs. These efforts will provide key information for scaling data visualization tools and platforms for larger-scale agricultural SC atlas resources that will need to be developed to visualize the wealth of data to be generated in the coming years. Pitfalls and Limitations: If data ingestion is problematic, we can either reduce metadata to a minimum description or establish a "stripped down" version of the HCA DCP. One limitation to our plan is that we are focused on scRNAseq, although other SC methods such as snATACseq, scMethylseq, as well as multi-omic methods are available and have been reported by the Ag community (7). If time permits, we will extend our approach to snATACseq data and analyses.

2. Furthering the aims of the AG2PI

The project brings together the crop and livestock communities at the very early stage of creating/using SC data. Thus now is a very opportune time to coordinate such a discussion with the HCA community who have built a world-leading SC data infrastructure. It is crucial that US and EU SC research and data infrastructure developments are aligned. To this end, the proposal includes key representatives of research and infrastructure from US and EU HCA DCP teams, USDA funded researchers, FAANG/ EuroFAANG members, and, through these persons, connections to key related groups such as AgBioData and Elixir. Collaboration with James Koltes (AgBioData advisory member; NRSP8 livestock bioinformatics co-coordinator, see letter) is providing continued helpful feedback from his AG2PI data reuse project.

We note that many interactions across groups, including with the HCA, occurs through the early-career data scientist, Kapoor. These interactions will benefit both her education and future career prospects. This is an important aspect of the proposal and a long-term aim of AG2PI.

We will test the utility of the current HCA DCP for Ag data storage, sharing and analysis. Using the seed grant outcomes, our Working Group will develop a proposal for a sustainable SC Ag data infrastructure; such preparation for future funding is also an aim of AG2PI.

Success will be measured in several ways. First, if we establish data standards that can leverage the HCA DCP, we should see an increase in using the HCA DCP, and thus in the number of FAIR Ag SC datasets. Second, our white paper and website on available tools of value to the Ag community are specific deliverables. Third, we should see SC data become visible on genome browsers, adding to overall genome annotations for G2P research.

3. Expected outcomes & deliverables

The primary outcome will be to lower the currently high bar to visualize and further use SC data, catalyzing substantial increases in SC data sharing and storage (RFP Priority area #1), and encouraging development of tools for SC re-use in G2P research in the Ag community (RFP Priority area #2). SC analysis is very powerful for measuring cellular phenotypes, even within novel samples such as organoids in phenotyping approaches to decrease animal use. An example of this is provided by collaborator Elisabetta Giuffra (see letter).

The following are anticipated outcomes and deliverables:
Aim 1: Demonstration that current crop and livestock metadata can be used to ingest scRNAseq data at HCA DCP and that tools within Terra (HCA DCP) can be used to analyze ingested data. A white paper and website summarizing tools for scRNAseq data exploration and analysis.
Aim 2: Initial tools for displaying scRNAseq data on genome browsers and querying outputs of scRNAseq analysis (e.g. gene counts, clusters/cell types) along with metadata in a web-accessible database.

Across Aims, we will organize a Working Group meeting to create the seed grant report and to develop plans implementing these outcomes through future USDA AFRI AG2PI funding.

4. Qualifications of the project team (0.5 pages) (Due to space limitations, only co-PIs are described. Collaborators describe their qualifications in their letters).
Christopher Tuggle-PI: Dr. Tuggle provides expertise in functional genomics and bioinformatics. He provides international leadership in genomics, including as founding co-Chair of the Functional Annotation of Animal Genomes (FAANG) consortium. He is also the USDA-National Swine Genome Coordinator, and a co-PI of the USDA-funded AG2P initiative.
Christine Elsik-coPI: Dr. Elsik provides expertise in genome informatics resources for species on importance to agriculture, leading projects such as the Hymenoptera Genome Database, MaizeMine, and FAANGMine, which use InterMine data mining warehouses and JBrowse genome browsers.
Peter Harrison-coPI: Dr. Harrison provides expertise in livestock and plant genomics, metadata standards and validation, data portals and data reuse. Dr. Harrison leads the FAANG Data Coordination Centre, chairs the EuroFAANG steering group and co-chairs the FAANG Metadata and Data Standards committee.
Nicholas Provart-co-PI: The Provart Lab's Bio-Analytic Resource (BAR) at bar.utoronto.ca, comprises tools for genome analyses in Arabidopsis and other plants, and receives 4M page views a month. He is a founding member of the International Arabidopsis Informatics Consortium, and is president of the Multinational Arabidopsis Steering Committee.

5. Proposal timeline (0.5 pages)

Category	Task Description	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Overall	Monthy meetings on project progress	x	x	x	x	x	x	x	x	x	x
	Hold WG meeting								xx		
	Write AG2PI report and White paper									xx	xx
Aim 1	Select data sets and metadata files	xx									
	Test data ingestion	xx	xx	xx							
	Tool information gathering			xx	xx						
	Testing tools in Terra HCA-DCP				xx	xx	xx	xx			
	Organize tool survey on-line display							xx	xx	xx	
Aim 2	Complete subcontract paperwork	xx									
	Tools for display of sc data on genome browsers	xx	xx	xx	xx	xx	xx	xx			

6. Engaging AG2P scientific communities & underrepresented groups
The collaborative group has gender-, community-, and national diversity, and includes scientists from academia and government working in plant and animal genomics, computational biology, and data science. All results will be reported at public meetings and on-line to engage a wide audience, and comments for improvement will be solicited from all communities.

Bibliography/References cited

1. Clark EL, Archibald AL, Daetwyler HD, Groenen MAM, Harrison PW, Houston RD, Kühn C, Lien S, Macqueen DJ, Reecy JM, et al. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biol* (2020) **21**:285. doi:10.1186/s13059-020-02197-8

2. Denyer T, Timmermans MCP. Crafting a blueprint for single-cell RNA sequencing. *Trends Plant Sci* (2022) **27**:92–103. doi:10.1016/j.tplants.2021.08.016

3. van der Wijst M, de Vries DH, Groot HE, Trynka G, Hon CC, Bonder MJ, Stegle O, Nawijn MC, Idaghdour Y, van der Harst P, et al. The single-cell eQTLGen consortium. *Elife* (2020) **9**: doi:10.7554/eLife.52155

4. Jha SG, Borowsky AT, Cole BJ, Fahlgren N, Farmer A, Huang S-SC, Karia P, Libault M, Provart NJ, Rice SL, et al. Vision, challenges and opportunities for a Plant Cell Atlas. *Elife* (2021) **10**: doi:10.7554/eLife.66877

5. Cole B, Bergmann D, Blaby-Haas CE, Blaby IK, Bouchard KE, Brady SM, Ciobanu D, Coleman-Derr D, Leiboff S, Mortimer JC, et al. Plant single-cell solutions for energy and the environment. *Commun Biol* (2021) **4**:962. doi:10.1038/s42003-021-02477-4

6. Herrera-Uribe J, Wiarda JE, Sivasankaran SK, Daharsh L, Liu H, Byrne KA, Smith TPL, Lunney JK, Loving CL, Tuggle CK. Reference transcriptomes of porcine peripheral immune cells created through bulk and single-cell RNA sequencing. *bioRxiv* (2021)2021.04.02.438107. doi:10.1101/2021.04.02.438107

7. Farmer A, Thibivilliers S, Ryu KH, Schiefelbein J, Libault M. Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. *Mol Plant* (2021) **14**:372–383. doi:10.1016/j.molp.2021.01.001

8. Becker D, Weikard R, Hadlich F, Kühn C. Single-cell RNA sequencing of freshly isolated bovine milk cells and cultured primary mammary epithelial cells. *Sci data* (2021) **8**:177. doi:10.1038/s41597-021-00972-1

9. Nayak R, Hasija Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics* (2021) **113**:606–619. doi:https://doi.org/10.1016/j.ygeno.2021.01.007

10. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* (2016) **17**:66. doi:10.1186/s13059-016-0924-1

11. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* (2021) **49**:D1046–D1057. doi:10.1093/nar/gkaa1070

12. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* (2012) **28**:3163–3165. doi:10.1093/bioinformatics/bts577

Scope of work

Please fill in the below table with the requested information. For example, please replace "PI" with the name of the person who is listed as the first investigator on this project; then repeat for each column with the corresponding names. You may add columns if your project includes more than four investigators.

Activity/Output	***Chris Tuggle***	***Peter Harrison***	***Chris Elsik***	***Nicholas Provart***
Overall grant management	**x**			
Aim 1- Data ingesting	**x**	**x**		**x**
Aim 1- Data analysis	**x**			**x**
Aim 2 Genome annotation	**x**		**x**	